ISCR

Institute for Scientific Computing Research

Laboratory Directed Research and Development Project Research Summaries

CASC Center for Applied Scientific Computing

SAVAnTS: Scalable Algorithms for Visualization and Analysis of Terascale Science

Mark Duchaineau

Center for Applied Scientific Computing

Summary:

Multi-physics simulation codes on LLNL supercomputers that are vital components of Stockpile Stewardship and other programs may produce dozens of terabytes of data during a given run. Great strides are being made to increase the efficiency and accuracy of the codes by harnessing thousands to tens of thousands of processors using scalable algorithms, but the efficient and accurate post-computation data handling and interactive exploration must also scale efficiently to reach the goal of a productive terascale simulation capability. Our goal is to offer multiresolution data selection and compression algorithms, coupled with optimal data access during interaction, to provide thousand-fold greater interactive performance and hundred-fold more efficient storage than current best practice. Success is measured by the reduction in disk usage, and the increase in interaction rates, with the effectiveness remaining constant or even improving as problem size and machine size grow exponentially.

Our focus is on devising new wavelet transforms and other hierarchies for compression and accelerated access and display of volumetric field data and boundary or contour surfaces. The central principle is output-sensitive calculations: storing and computing exactly enough information given the actual usage patterns of the scientist who is looking at the simulation results. We find that the useful information content of a 3D field such a pressure or density is quite sparse and readily compressed after the application of an appropriate wavelet transform. This results from the property that wavelets tend to automatically find and exploit coherence in both space/time and frequency/scale. Wavelets are well understood for regularly spaced grids filling a complete tensor-product brick of space. The innovations required for large-scale laboratory applications include the extension to highly adaptive or unstructured settings, and to arbitrary surfaces that typically can not be represented as anything resembling a regularly spaced grid. More fundamentally, the work performed should be proportional to the sparse post-transform information content at as many stages as possible of the end-to-end data flow going from simulation to scientist. This leads to a suite of connected optimization problems that we address.

During FY2000 we devised a new type of wavelet for general surfaces that is the first to have bicubic precision while allowing sparse, local transforms using small filters at all steps in the transform. Thus our methods are the first practical high-precision wavelets for use by large-scale simulations for surface storage. Supporting the wavelet transform are new conversion procedures we have devised that automatically shrink-wrap to a semi-regular form the arbitrary-shaped geometry that contains any number of topological handles and connected components. An example is a record-breaking simulation of a Richtmyer-Meshkov instability forming in a shock-tube experiment (for which two members of our LDRD team were co-recipients of several awards, including the Gordon Bell Prize in the Performance Category at the IEEE SuperComputing conference in November, 1999). The complete surface if represented conventionally contains 460 million triangles and occupies 13GB of disk space. The shrink-wrap results depicted allow application of our wavelet transform, which our experiments indicate is likely to reduce the storage to under 260MB, a fifty-fold space reduction. The wavelet transform and coding are at least ten times faster than unstructured surface-mesh compression schemes and previous wavelet surface compression methods based on infinite-width filters.

In FY2000, we plan to increase performance on per-view surface optimization time by two or more orders of magnitude using pre-optimization on subregions, and complete a working prototype for the wavelet surface compression.

Sapphire: Scalable Pattern Recognition for Large-scale Scientific Data Mining

C. Kamath, E. Cantu-Paz, I. Fodor, and N. Tang

Center for Applied Scientific Computing

Summary:

There is a rapidly widening gap between our ability to collect data and our ability to explore, analyze, and understand the data. As a result, useful information is overlooked, and the potential benefits of increased computational and data gathering capabilities are only partially realized. This problem of data overload is becoming a serious impediment to scientific advancement in areas as diverse as counter-proliferation, the Accelerated Strategic Computing Initiative (ASCI), astrophysics, computer security, and climate modeling, where vast amounts of data are collected through observations or simulations. To improve the way in which scientists extract useful information from their data, we are developing a new generation of tools and techniques based on data mining.

Data mining is the semi-automated discovery of patterns, associations, anomalies, and statistically significant structures in data. It consists of two steps: in data pre-processing, we extract high-level features from the data, and in pattern recognition, we use the features to identify and characterize patterns in the data. In this project, we are developing scalable algorithms for the pattern recognition task of classification. Our goal is to improve the performance of these algorithms, without sacrificing accuracy. We are demonstrating these techniques using an astronomy application, namely the detection of radio-emitting galaxies with a bent-double morphology in the FIRST survey.

In FY2000, we focused on three tasks: (a) improving the performance of decision tree algorithms; (b) identifying bent-double galaxies in the FIRST survey; and (c) incorporating our research into software to make it easily accessible to LLNL scientists. In decision trees, we considered oblique trees, where a decision at a node uses a linear combination of features, instead of a single feature. As this is essentially a search in a high dimensional space, we investigated the use of evolutionary algorithms to solve the optimization problem. Our research showed that combining evolutionary algorithms with decision trees resulted in better and faster classifiers. On a data set with 50 features or dimensions, one of our new algorithms, Oblique-ES, was more accurate (79 % vs. 73 %) and four times faster than the current best oblique classifier. Another algorithm, Oblique-GA, which gave the most accurate results (85 %), was twice as fast. In contrast, the traditional axis-parallel tree, while fast, resulted in accuracy (58 %) that was just a bit better than making a random decision.

For the bent-double problem, we focused on galaxies composed of three blobs. Using features from the FIRST catalog, we completed six inner iterations of data mining, improving the features extracted. We completed two outer iterations, increasing the size of the training set from 195 to 495 examples. In the process, we reduced the classification error by 50 percent.

In FY2000, we released the Alpha version of our software, which includes serial, object-oriented versions of decision trees, wavelets and wavelet de-noising techniques, evolutionary algorithms, and support for a parallel infrastructure. In addition, we published 12 papers, 8 conference and workshop presentations, and 3 records of inventions. We co-organized two data mining workshops and actively participated in university collaborations.

In FY2001, we will complete our work on parallel scalable decision tree algorithms and software, complete the detection of bent-double galaxies using the techniques developed, and focus on the use of evolutionary algorithms to improve the performance of neural network algorithms for classification.